WHAT ABOUT KENTUCKY'S TEST?

THE QUESTIONS

Kentucky parents and teachers ask a lot of questions about Kentucky's testing program. Some of the basic questions are: What does it cover? How is the test built? Who builds it? Is it valid? What is a standards-based test? Why does it take so long to get the results? Can an essay really be graded consistently? Why aren't students held accountable? These are just a few of the questions asked that we will consider.

GENERAL MATTERS

What is tested and when is it tested?

The first time a student meets the Kentucky testing system, which is called the Commonwealth Accountability and Testing System (CATS for short), is in the third grade. In April third graders take a multiple-choice test called the Comprehensive Test of Basic Skills (CTBS/5), which is produced by the CTB/'McGraw-Hill Corporation. Because this test is used nationwide, Kentucky students can be compared to students in other states. This test is repeated in grades six and nine.

In grade four, students write parts of the Kentucky Core Content Test (KCCT for short) for the first time. This test is very different from the CTBS/5. First, the students write essay type answers (called open-response), as well as multiple-choice. The open-response answers are limited to one page. A second difference is that the test is limited to the Core Content, which is specifically what Kentucky students are expected to know and do at the fourth grade level (more about that later). The test asks questions about reading, writing and science. For reading and science there are six open-response questions that count for the student, and one that is being evaluated for future tests. This one does not count in the student score. The writing test offers the student two questions, but they only have to answer one. In addition, fourth graders produce a collection of expanded work, representing their best efforts, called a Writing Portfolio (more about that later, too).

Grade five students continue the KCCT, but in different subjects than fourth grade. While mathematics and social studies are tested in grade five, the format of this test resembles the fourth grade reading and science test with six "live" open-response questions and one experimental question. Two new subjects are also tested for the first time: arts & humanities, and practical living/vocational studies. These tests are shorter, having two open-response items and one experimental item, along with fewer multiple-choice than what is on the mathematics or social studies part. The following bullets summarize testing in middle and high school, which is very parallel to the testing in elementary schools mentioned above:

- Students in the sixth grade repeat an age appropriate version of the CTBS/5.

- Seventh graders write tests in the same subjects as fourth grade, with the same number of questions at a grade appropriate difficulty level.

- Eighth grade repeats the same subjects as fifth grade.

- Ninth grade repeats the CTBS/5.

- Tenth grade takes the KCCT in reading and practical living/vocational studies. These tests have the same number of questions as these subjects had in earlier grades, but the questions have increased in difficulty at each level.

- Eleventh grade is the most heavily tested grade in high school. Students write the KCCT in mathematics, science, social studies, and arts & humanities.

- Since many students graduate at the end of the first semester of grade twelve, only two parts of the CATS are completed: in twelfth grade, the writing portfolio which can be finished the first semester, although it is not due until April, and the writing question (called writing on-demand) which is also administered in April.

One of the strong points of CATS is that it does not depend on a single type of testing. The KCCT includes multiple-choice in every subject and grade from three through eleven, open-response in grades four, five, seven, eight, ten, and eleven, writing questions and portfolios in four, seven and twelve. The variety of testing methods allows students to show a greater range of their abilities.

How does Kentucky's testing system compare with other states?

Kentucky has an advantage over some of the largest states where the cost of scoring open-response questions forces them to use all multiple-choice, or only a few open-response in a limited number of subjects. Kentucky has an advantage over some of the smallest states, which have inadequate resources to construct a test that meets their own needs expressed in something like our Core Content.

Another advantage of CATS is that Kentucky tests more subjects than the states that limit testing to reading and mathematics, and sometimes writing. In addition to testing a larger number of subjects, Kentucky tests at as many or more grade levels than some other states. Some writers have pointed out that states that only give a single multiple-choice test in reading and math, but give it every year, have a better picture of how the individual student is keeping up nationally. In a limited sense this is true, but the test may not match very well what is taught in the state, and the breadth and richness of a curriculum like Kentucky's is lost. Such tests are much more susceptible to cheating and teaching to the test rather than teaching to a body of knowledge.

There are also differences at the high school level between Kentucky and some of the other states. Some states test at the end of specific courses, such as algebra, or U.S. history, or American literature, instead of the general mathematics or social studies tests that Kentucky uses. These "exit" exams are used for a variety of purposes from assigning final grades to entrance to a following course. The goal is to make the student more responsible, but a one-day test may not be as descriptive of a student as a semester or year of daily work.

Some states use an exam to decide whether a student is ready to graduate. Kentucky's system is devoted to improving instruction, not to testing individual students. This is an important enough issue that we will consider it further.

Why doesn't Kentucky use tests for promotion and graduation?

Some states have what are called promotion or "exit" exams. Logically, it seems that this would put more pressure on students to do their best in order to pass to the next grade. They would be more "accountable." If the world were simple and completely logical maybe this would work, but in the real world there are some surprises hidden in such testing. For example:

- Promotion tests almost always increase the number of students held back (retained) in the prior grade, resulting in increased costs for a few years until new school population patterns are established.

- A second result is that the increased retentions lead to increased dropouts at grades eight through twelve, especially the grade prior to the administration of the test. This is the opposite of what Kentucky has been seeking to do with regard to dropouts.

- Another problem is that if the test is modified to keep approximately the same pass rate, then it does not seem to measure as much or require as much educational achievement.

So, in order to have high pass rates and a hard test, just have the teachers teach better. This brings us back to the beginning. Retaining students, according to many studies, does not motivate students to perform better, but better teaching and smaller classrooms at the primary grades do.

Kentucky's testing system aims to make very clear what is to be taught, and how good performance should look, so that teachers know what will be tested and at what difficulty level they must teach the content. Student motivation issues tend to be reduced in well-taught classrooms, but there will always be the one student in twenty who cannot be externally motivated, even by a high stakes test.

Kentucky has chosen to not ask of a standardized test something that it cannot do: that is, give a picture of the development of a child from a one-day paper and pencil test.

Kentucky has sought to put data together at a level where it can give an accurate picture, which is at the school level.

Why doesn't Kentucky hold students accountable?

This was partially discussed above under the promotion/graduation questions. Kentucky has considered a number of proposals to increase student accountability. The problem is strongest at the high school level, begins to appear at the middle level, and is less of an issue at the elementary level. Including KCCT performance as a small part of the student GPA is an example of a proposal that was considered and not accepted. Since the proposal was optional, most high schools indicated unwillingness to engage in the extra calculations this would involve. Other proposals have surfaced, both locally and nationally, but no really effective means of motivating low performing students, other than the classroom teacher, has been found. Kentucky will not adopt student accountability until a successful method has been found.

TEST CONSTRUCTION

Who writes the questions for the Kentucky test?

Unlike standardized tests, which may be built in other states like New Jersey or California, and may or may not be related closely to state standards, the Kentucky test is related to nationally recognized standards as well as Kentucky standards. With the exception of the CTBS/5 component, Kentuckians create the Kentucky test. The writers are Kentucky teachers who are experts in the subject area for which they write questions. They are among the best Kentucky teachers who have exhibited expertise in teaching, have shown the ability to teach by various methods to meet the wide range of student needs, and have come to the attention of their principal and/or District Assessment Coordinator who recommend them. They are assigned to a Content Advisory Committee (CAC), which meets in the spring of the year to write questions.

The CAC may write as many as thirty open-response questions, and as many as 100 multiple-choice questions in each subject area. The CAC participants write with three objectives in mind: to improve the quality of questions used on the test, to make sure all parts of the Core Content are covered, and to provide replacements for questions. Approximately 20% of the questions are replaced each year.

The questions are then submitted to one of our contractors, currently WestEd, which is a California company with expertise in question writing and building tests. They edit the questions and balance the wrong answers (the distractors) so they are not correct, but not so ridiculous that no student would choose them. In the fall of the year the CAC comes together again and looks at the revised questions and makes selections for testing. As you can see, this means that Kentucky teachers write the questions for the test and pick out which ones will be used each year.

Who puts Kentucky's test together?

The contractor builds six different forms of the test in each subject area. The multiple forms allow the full coverage of the Core Content in that subject, which is important for evaluating a school. Each form has one experimental question that the student answers, but the form is labeled A or B which allows the testing of two experimental questions. The live questions remain the same on both forms. Several statistical measures of the quality of the question are accumulated as it is tested and used, such as percentages of students who select each answer on multiple-choice questions, p-values (a measure of difficulty), bi-serial correlations, and others. More information about these statistical matters is available in the *KCCT 2000 Technical Report* which is available from the Office of Assessment and Accountability, Kentucky Department of Education, Frankfort, Kentucky and on the website for the department. These statistical tools are also used to make the forms as comparable in difficulty as possible.

One form of the test is selected for use with visually impaired students. That form will be translated into Braille, produced in large print, and recorded on tape so a student can play it, and back it up, to hear the questions a second or third time if necessary. Some impaired students may answer on a computer. The intention is to give the impaired student the same chance that other students have to answer the questions successfully.

Another form is selected for "scaling" and "linking" the test from year to year. That form is held stable from one year to the next so that changes in performance can be measured as real changes. When the other forms are "scaled" within the year to the linking form from year to year, the gains exhibited are genuine. A pattern for selecting the linking form has been developed so that one form is not held stable for several years, which leads to "aging," or the questions becoming familiar and known to teachers, which would distort the results.

What guides the content of Kentucky's test?

The Core Content has already been mentioned several times. What people usually mean by a standardized (norm-referenced) test is one that is connected to the expected performance of a normally distributed group of students at a particular level in school. A representative sample of students set what is presumed to be normal performance on the test. The sample sets the national mean, quartiles and percentiles, which is the way scores are usually reported.

A standards-based test is tied to a set of statements about what students should know and activities they should be able to do. The statements are fixed, and the distribution of the students will not follow a normal curve in most cases. The boundaries between categories are called cut-points. No matter how many students move into a higher category, the boundaries or standards do not change. Kentucky divides students into four categories: novice, apprentice, proficient and distinguished. The student who does nothing is categorized as novice non-performing. The goal is for all schools to be

proficient by 2014. Proficient is defined as a score of 100 on a 140-point scale. For the school to achieve that goal nearly all students must also move to proficient.

The *Core Content for Assessment* is a document that states the minimum that students must know and do in terms of what will be tested. This document is available on the Kentucky Department of Education website. Students learn much in school that cannot be tested, but whatever teachers choose to teach must include the Core Content in their subject area. If the course is Algebra, many concepts will be taught that cannot be included on the KCCT, but certainly the teachers must make sure that students learn the particular algebraic concepts that are mentioned in the Core Content, because they will be tested.

Just what is in the Core Content?

The Core Content describes what to know and do at three levels: elementary middle and high school. Seven subjects are included in the Core Content: reading, mathematics, science, social studies, arts & humanities, practical living/vocational studies, and writing. Each content subject is divided into subdomains: mathematics, for example, has four, which include number/computation, geometry/measurement, probability/statistics, and algebraic Ideas. Science has three subdomains: physical science, earth and space science, and life science. Other subject areas are similarly organized.

The next division of the content is that each subdomain is divided into sections. For example, the 4th grade science subdomain of earth and space science is sectioned into properties of earth materials, objects in the sky, and changes to earth and sky. The final layer in the content is the specific statement of the content under the subdomain and section. These are called “bullets.” One bullet under properties of earth materials for the 4th grade says, “Earth materials include solid rocks, and soils, water and the gases of the atmosphere. Minerals that make up rocks have properties of color, texture, and hardness. Soils have properties of color, texture, the capacity to retain water, and the ability to support plant growth. Water on Earth and in the atmosphere can be a solid, liquid or gas.” The teacher is told the broad topics to teach, but not how to teach it.

Who created the Core Content?

Once again, Kentucky teachers, the experts in their fields, wrote the Core Content. The committees of teachers who did this task consulted and considered what national organizations had published. For example, the National Council of Teachers of Mathematics has extensively documented content at each grade level in ten strands. These national content standards were considered in Kentucky’s Core Content writing. Similar standards exist in language arts, science and social studies. The teachers in arts and humanities, and practical living/vocational studies had less guidance from national organizations.

Has the Core Content narrowed or dumbed down what students have to know?

One of the most common accusations leveled at any statewide testing program is that teachers teach to the test and dumb down the curriculum. There are several levels of teaching to a test. The first is obtaining the test questions and drilling students over correct answers to the test. This clearly is cheating, artificially inflates student scores, and contributes very little to student learning. Kentucky seeks to avoid this kind of teaching by keeping the questions secure, having teachers sign non-disclosure statements, making it inappropriate to copy down the test questions, or even making a list of topics covered. This is actually not necessary since the Core Content is the list, and the six forms cover all or nearly all in a given year.

At another level, however, Kentucky does encourage teaching to the test. Since the test and the Core Content match so closely, every bit of the Core Content needs to be taught, sometimes in multiple ways. This procedure assures that students can answer whatever question comes up on that topic. In some years, questions that have been used on previous CATS tests, and that will not be used again, are released so that teachers have examples of what students have to do to succeed. Examples of student papers at the four performance levels are also released (without names of course). On the other hand, classroom topics need not be limited to the Core Content, however, one of the primary reasons schools are not successful on the CATS is that they do not teach the Core Content. This failure to address the Core Content has been revealed by the school auditing process, which has been conducted in recent years. The most successful schools have rich and varied curricula, but do thoroughly cover the Core Content.

What are standards?

There are several kinds of standards. The Core Content, already mentioned, is one type of standard, a content standard. Every child is supposed to be able to know what the Core Content specifies, and do the skills described at the appropriate grade level of difficulty. The content standard that Kentucky uses is certainly not everything a person should know, but it is the minimum that a person must know to be considered educated and able to function in society.

A second kind of standard is a performance standard. This is a boundary mark that is the target for a student to achieve in order to be classified a certain way. In the high jump, for example, a jumper in high school who exceeds six feet six inches would be considered proficient, in college it would take a jump of six feet ten inches to be considered proficient, and a world class jumper might have to reach seven feet and a few inches to be considered proficient. These are benchmarks that indicate whether the person is going to be competitive. In education the concept is the similar. There are certain scores on a test that are benchmarks. They are called cut scores. Everyone who reaches the first cut score in Kentucky is considered an apprentice. Those below

that first mark are novice. Those above the second cut score are considered proficient, and those beyond the third are considered distinguished.

What is "standards setting?"

Standard Setting is the process of deciding where the boundaries are between the four categories that Kentucky uses in describing student accomplishment. Standards were set in 1992 for the old KIRIS test by a relatively small group of teachers. While those standards generally worked well, there were problems in some subjects in that it was difficult for students to actually show the higher performance categories. When the KIRIS was revised into CATS in 1998, it was clearly necessary to set new standards for the new test. This was done during the timeframe from late 1999 to early 2001.

Approximately 1600 teachers participated in a six-step process designed by the Kentucky Department of Education (KDE) and a panel of six national testing experts. Three different methods of setting standards were used, two of which did not even exist when Kentucky first set standards in 1992. The methods used student work, teacher evaluations of classroom performance, and difficulty rankings of actual test items to set the standards. A final step synthesized the varying results from the three methods into Kentucky's standards that are hoped to be stable for many years. Contrary to some critic's claim that the new standards turned CATS into a norm-referenced test, this is not the case. The new cut-points or standards are clearly tied to the Core Content and to specific points that students must achieve, regardless of the percentage of students that achieve that category. An additional result of the new standards was the creation of a set of clear definitions of what each performance level represents, definitions useful to both teachers and parents.

What is the difference between a standards-based test and norm-referenced test?

As indicated above a standards-based test expects students to reach a certain level on the test to reach a category. It does not matter how many students achieve the standard. The mark remains the same. In Kentucky, at present, more students are in the bottom two categories than are in the top two. The goal is to reverse that situation by 2014. For a norm-referenced test, students are assumed to follow a certain curve with 68% of the students within one standard deviation on either side of the mean, and approximately 95% within two standard deviations on either side of the mean. If students begin to increase their scores the test has to be re-normed to remain useful. That means the target for the student moves as scores improve, whereas the target for the student remains stable, and therefore known to all, in a standards-based test.

Who gives the test?

Once the forms have been constructed, they are shipped to a second contractor, currently Data Recognition Corporation of Maple Grove (Minneapolis), Minnesota. There the test booklets, and answer booklets are printed, quality checked, boxed by school and shipped to the 176 school District Assessment Coordinators. These

administrators at the local level check the boxes to make sure each school has adequate materials, and distribute the tests to the schools a few days before the testing window (around late April and early May). The school has a Building Assessment Coordinator who is responsible for making sure that teachers who administer the test follow instructions. Some students may take the entire test over several subjects in one location. Others may be in a different room each day. Schools have several different patterns they may follow regarding how much testing is done each day. The crucial issue is that all students at a grade level must do the same sections of the test on the same day. The Kentucky Department of Education provides Administration Manuals for teachers to use that tell them exactly how to give the test and exactly what to say, so that all students have an equal chance to do well.

How do we know the test was given fairly?

The main safeguard of fairness is the integrity of Kentucky teachers. While we sometimes read in the papers about a teacher doing something illegal, the fact is that teachers are among the most honest and truthful groups of people in the state. Even if you have had a bad experience with a teacher, that does not necessarily mean they are dishonest or untruthful. In addition, there is an "allegation" process where parents, teachers, or administrators can file a complaint or an admission if something was done incorrectly. A division of KDE that is completely separate from the Office of Assessment and Accountability investigates the allegations. If the allegation proves true, it may fall into one of two categories. One includes those incidents that do not affect student scores. The other category includes those allegations that do affect student scores. Student scores may be changed to zero, which punishes the school for not administering the test appropriately. It should be noted that in these cases, parents still receive a score report for their children that has original scores, but a zero score is used for purposes of school accountability at the school level.

The number of allegations per year ranges from 100 to 200. In light of the more than 30,000 teachers who administer tests each year, this is a very small amount. Test scores change for a few hundred students each year of about 400,000 tested each year.

What about portfolios?

Kentucky is one of the few states that have a statewide portfolio requirement that is used to aid in evaluating schools. The submission of writing portfolios occurs in grades four, seven and twelve. Work on the pieces submitted may take place at any grade level. Students submit a specified number of pieces that exhibit ability to complete different kinds of writing like personal narratives, persuasive, or practical workplace writing. One piece must be from a subject other than language arts. The portfolios are scored at the school according to specific requirements (called rubrics) by groups of teachers, language arts teachers at some schools, and all teachers at others. Each year KDE in cooperation with a contractor conducts an audit of 100 schools: fifty

selected randomly, and 50 selected because they exhibited a large change in scores. The accuracy of scoring is verified for these schools.

What is an alternate portfolio?

In a prior question we mentioned steps taken to make it possible for students that are impaired to have an equal chance to perform well. There are, however, some students so severely impaired intellectually or physically that they are unable to perform with a paper or pencil test. In Kentucky, somewhat less than one percent of the students fit in this category. A special means has been developed to measure the progress of these students, called the alternate portfolio.

Each student with a severe impairment is in a classroom with fewer students, although they may spend some of their day in a regular classroom, with modified assignments, and with the help of a supporting person. They have an individualized plan of educational goals, which are selected from Kentucky's Academic Expectations and the *Core Content for Assessment.*

The Alternate Portfolio is the tool for assessing progress toward the goals selected for the student. The required contents of the portfolio include a table of contents, a student letter to the reviewer, a parent letter validating the portfolio, the student's schedule, a summary of job exploration at grade 8 or a resume at grade12, and five entries which represent the required subject areas at the student's grade level.

Alternate portfolios are scored by two teams of two teachers who are familiar with the construction of alternate portfolios. Scoring takes place at the regional level. Agreement between the teams makes the score final. Disagreement leads to scoring by a state expert whose decision is final. A single category (novice, apprentice, proficient or distinguished) is given to the portfolio. In the accountability index for the school, the student score counts in each subject area required at that grade level.

Validity and Reliability

Does the test measure what it is supposed to?

Validity is the appropriateness, meaningfulness and usefulness of the conclusions drawn from test scores. KDE takes very seriously the standards of national professional organizations relating to validity. Careful data is maintained about both the teachers who write the questions and about the match between the Core Content and the KCCT. When the teachers write the questions, they assign a primary and possibly a secondary Core Content "bullet" that the question is intended to measure. The contractor's experts evaluate these assignments and give feedback to the Content Advisory Committee if they disagree. The question is then reconsidered by the CAC. Problematic questions may never make it to the test, but if the question is regarded as exceptional, it may be tested. Research concerning how students answered each experimental question (i.e.,

each pre-test question) may well enlighten the CAC regarding whether these questions allowed the desired response from students.

A second means of checking whether the test measures what it is supposed to is an annual report of all the assigned Core Content codes on the test. This report is used to see if the test is properly balanced and covers all the content bullets in the six forms in a subject area. This report is compared to a document called the *KCCT Blueprint* that specifies the percentage of questions on the test for each subdomain. Kentucky teachers created the Blueprint, with the help of KDE staff. The annual report indicates whether the percentages specified in the Blueprint are being met. The report guides the writing of questions to specific topics where there may be a gap, and also guides the form building process.

One of the most important issues with regard to any test is whether the test considers appropriate criteria such as cognitive complexity (how hard the questions causes a person to think) or content quality (how well the question measures the content). One way of looking at this question is whether the student who answers the questions can show proficient and distinguished performance. This issue has been carefully addressed by the CACs. The setting of new standards also has an impact upon this question (See below). Another way of answering this question is whether high scoring schools do things differently than low scoring schools. Several studies have accumulated, as well as results from school audits, that indicate that high scoring schools are very intentional in aligning their curriculum to the Core Content, have rich and rigorous curricula, and have aligned classroom assessment with the types of assessment that appear on the KCCT.

Does the test measure reliably?

Since many who ask this question are referring more to whether the test is accurately scored, than to formal reliability, we will consider that separately below. With regard to the reliability of the KCCT, in reading, mathematics, science and social studies at most levels, the reliabilities are between .80 and .89 which is excellent. For the shorter tests in arts & humanities and practical living/vocational studies, as expected the reliabilities are lower. They were .60 to .69, which is acceptable. For more information concerning reliability see the CATS 2000 Technical Report.

Is Kentucky's test fair to all students?

We have mentioned some fairness issues in earlier questions. We briefly discussed some of the means of allowing impaired students to have an equal chance to succeed (these are called accommodations). We also considered fairness in administering the test. The most frequent fairness concerns involve gender and race. There is a consistent pattern over the years of girls outperforming boys in language arts and social studies at the middle and high school levels. There is a second pattern of boys performing better than girls in mathematics at the high school level. There is a consistent pattern in the test results of those with an Asian heritage outperforming all

students, and of Caucasians outperforming African-Americans and Hispanics. Do these results reflect bias in the test or are they an accurate reflection of the results of the educational process?

Kentucky uses two methods to make sure that such performance differences are not due to bias in the test. The first means is the Bias Review Committee (BRC). This group, which represents a broad cross-section of educators, business people, and special concern groups, meets twice annually. In the spring this group reviews reading passages that will be used by the CACs to write questions. The BRC looks for concepts that are only known to a few at the grade level, things that might offend or distract students from a racial, religious or social group, things that are outside the experience of a social grouping, or passages that do not lend themselves to use by the blind or hearing impaired. The fall meeting of the BRC is spent reading the actual questions that will be considered for experimentation for the same kinds of bias mentioned above.

The second method of finding bias is quantitative, that is, it is based on mathematical analysis. The method is called Differential Item Functioning (DIF). This method compares how the item worked in comparison to all the other items on the test of like kind. If an unusual pattern for an item is discovered between groups of students, it may be sent back to the BRC to be rechecked for bias, or the question may be removed from the test. Kentucky uses the most elaborate and complex method available for checking DIF, and possible bias.

If the test is not biased, then what explains the performance differences between groups? This becomes an instructional issue. Are girls **expected** to do as well as boys in mathematics? Are African-American students **expected** to do as well as Caucasian students. The Instructional Equity team of KDE, as well as the Division of Equity address equality of opportunity and of expectations. The issue of equal opportunity and equal expectations is also a component of the audit process for low performing schools.

Who checks to make sure things are done right?

We have already answered this question in part. The contractors and the Bias Review Committee check the work of the Content Advisory Committees. In turn, KDE checks the work of the contractors. In past years, teams from the Division of Assessment Implementation and the Division of Validation and Research visited each site where Kentucky tests are built, printed, scored and reported. Specific points of concern were identified in advance and carefully reviewed by KDE staff on these visits. In addition, there are many advisory groups that assist in making Kentucky's test one of the best in the land. A group of nationally recognized testing experts advise KDE on technical issues. This group, the National Technical Advisory Panel for Assessment and Accountability, or NTAPAA, meets quarterly. There are other in-state groups of advisors representing teachers, principals, superintendents, school boards, parents, professional groups, business people, chambers of commerce, the legislature, and others. New plans are passed before these groups before the Kentucky Board of Education (KBE) takes action. The KBE has the ultimate responsibility for making sure

things are done right. Despite what a few vocal critics might say, attendance at these meetings soon demonstrates that Kentuckians are dedicated to building both the best test possible, and an educational system that is successful.

SCORING

How do we know the test is graded fairly?

Kentucky assures fairness in scoring the Kentucky tests by contracting with independent contractors who have no vested interest in the outcome. The contractor is experienced in scoring and has many checks and rechecks built into the scoring system. As an example, the most experienced scorers reread 2% of all the questions to make sure that the original scorer is on track. The papers are randomly selected and the original scorer never knows which ones will be read. This is called a double read process. A second method is that scorers are organized into teams of ten with a leader. Once a day the leader reads approximately ten papers from each of his/her ten scorers. This represents 7 to 10% of each scorer's daily production. If a scorer has strayed they are immediately put back on track, and all the papers they scored that day may be re-scored. Kentucky requires an 80% perfect agreement rate for scorers to qualify to score Kentucky papers. This is the highest requirement of any of the states served by the current contractor.

Can open-response really be scored consistently?

While 80% perfect agreement between scorers doesn't sound very good initially, it becomes more impressive when we realize that it is higher agreement than is common for classroom essays. It is also easier to accept when we realize that the student does not suffer any consequences if his answer is scored incorrectly, if only one of several questions is off by one point. At the school level the questions that were incorrectly scored down are somewhat compensated by those incorrectly scored too high.

Who scores Kentucky's test?

The contractor hires those who score the tests. Kentucky requires that all scorers have at least two years of college, however, over 90% of the scorers have a college degree and many have advanced degrees, especially teachers and retirees. A sizable number of scorers at the six or seven sites that score Kentucky papers are teachers, but many other professions are represented as well. Teachers do not always make the best scorers, because some cannot accept the Kentucky rubric (scoring guide) without challenging it. This is important because KDE obviously wants the scoring to go according to the Kentucky designed and built rubric. The current contractor has scoring sites in Minnesota (at least five locations), Chicago, Cincinnati and Wilmington, North Carolina. The ethnic composition of scorers is approximately 13% minorities, which closely matches Kentucky's 15% minority student population. More females score than males, but then there are more female teachers in Kentucky schools.

What guides scoring?

The most significant piece of the process of accurate scoring, however, is the care with which the scoring guide (rubric) for a given question is written. The CAC member that writes the question also writes the scoring rubric. Using the rubric, they describe what student work will look like for each of the score points assigned. Most questions have either four or five total points possible, and the rubrics often specify how half points can be achieved. At the end of the rubric all scores are converted to a standard four-point scale with no half points. Kentucky rubrics have drawn praise from the contractor's readers for their completeness and ease of use. The experienced scorers used by the contractor become very capable of making consistent decisions about student papers hour after hour and day after day.

Reporting

Why does it take so long to get the results?

Scoring essay type questions for a whole state (well over 400,000 students) takes time. Just the packing at schools and unpacking at the contractors with the checking in of every paper takes several weeks. Two or three days are necessary to score the six questions on one form. There are six forms, and six subjects with the multiple forms, and the writing test, all of this at three different levels. So, the result is two months or more just to score. Then all the statistical work must be done to produce the information for individual students, the schools, the districts and the state. Even the simple printing and shipping of reports takes a great deal of time. So the time from the end of May to mid September turns out to be short in terms of producing what Kentucky needs. A simple multiple-choice test could give us quicker data, but a lot less information about how Kentucky schools are achieving.

What is an accountability system?

One of the most confusing aspects of Kentucky's testing system is the difference between the KCCT and CATS. The first is the actual test. The second is the accountability system that in fact includes the CTBS/5 and KCCT plus other indicators of school performance. The objective of CATS is to have the same goal for all schools, proficiency by 2014. But schools are starting at different points. Some schools are already excellent, but some are not.

CATS is designed to measure progress toward the goal. Simply put, a starting point was established for every school during the 1999 and 2000 biennium. The new standards were applied to the scores for those years to establish the starting point. Proficiency is defined as a score of 100 on a 140-point scale. A line is drawn from where the school was in 1999-2000 to a score of 100 in 2014. This creates a chart with

a line connecting two points, which is called the goal line. Schools whose score (or accountability index) is at or above the line are meeting the goal and are eligible for financial rewards.

In addition to the goal line, a second line is drawn from the 1999 and 2000 biennium point to an index or score of 80. This line is called the assistance line. Schools in between the two lines are "progressing" if their scores are increasing. These schools are eligible for smaller rewards. If the school scores remain the same or declines the school receives no rewards. Schools below the assistance line undergo a state review, and those in the bottom third of schools below the assistance line are audited to determine what financial and professional help they need to improve.

Because there is always the possibility of measurement error in any type of scoring (all tests have at least some measurement error), the goal line is actually drawn to a point slightly below 100 to take this possible error into account. Similarly, the assistance line is drawn to a point slightly below 80 to take possible error into account.

In addition to the above rewards system, for schools that are improving there are five recognition points where additional rewards may be earned. Also, the top five percent of schools, if they are above the fourth recognition point, may be designated Pacesetter schools.

How do rewards and assistance work?

Each year the Kentucky Board of Education determines the amount of money available for rewards. Information is gathered on the number of schools and the number of teachers in those schools in order to calculate the value of a share of rewards. Schools that are above their goal line, and have met their novice reduction and/or dropout goals, are eligible for three shares of rewards. Improving schools between the goal line and the assistance line receive one-half share of rewards. Schools that exceed for the first time one of five recognition points receive one share of rewards. "Pacesetter" schools that are past the fourth recognition point and that have not declined in the previous two biennia, and that are in the top 5% of schools and have met their novice reduction target, are entitled to a share of rewards. The dollar amount the school receives is the number of shares it is entitled to times the number of full- time teachers. School councils decide how to spend reward money and may choose from several options including materials, supplies or bonuses for teachers and other staff.

For schools below their assistance line, some or all of the following forms of assistance may be received: an invitation to draft a school improvement plan, a scholastic audit to recommend specific assistance needed, Commonwealth School Improvement Funds, a highly skilled educator to provide advice, and an evaluation of school personnel. The goal of assistance is to aid the school in beginning the process of making continuous improvement toward proficiency by 2014.

What is NAEP and why should I care?

NAEP means the National Assessment of Educational Progress, and is frequently called "the nation's report card." Tests are given to a sample of students every four years in a subject. Currently reading, mathematics, writing and science are tested. Kentucky has participated in this testing program since the beginning of state level testing in 1990. Results over the past decade show that Kentucky has been improving in the tested subjects in reading and math at the 8th grade level. Kentucky students are drawing near to the national average in both 4th and 8th grades. NAEP serves as a partial check that the progress made on the CATS is real and genuine. Other national programs like the ACT also provide some evidence. A growing number of high school students are taking the ACT. Normally when a larger number take the test, the assumption is made that scores will go down because a larger number of less able students are taking the test. The good news for Kentucky is that as the number of students taking the test has grown, scores have remained relatively the same.

How do I get the KCCT results for my child?

In the fall of each school year in mid-September each school receives the scores of students who participated in testing the previous spring. These scores for individual students are sent to parents or guardians a short time after the school receives them. The scores for CTBS/5 are received in mid-August, but are sent to parents after the beginning of the school year in most cases. If you do not receive scores for your child who is in the 4th through 12th grade call your school.

How do I learn about my child's school?

In addition to a report for the individual student, the school is also required to produce a document called the School Report Card. This document tells important information about the school, and the school's performance on the CATS, attendance rates, how many children were held back for a second year in a grade, how much money is spent per student, parent participation, the percentage of teachers with degrees in what they teach, the percentage who have a master's degree, and many other topics. A printed version of the School Report Card is sent home to parents by mid-January. The School Report Card for each school can also be viewed on the Kentucky Department of Education website.

Other Matters

Who does not have to take the test?

Less than one percent of Kentucky students are excused from the test. These students include:

- Students who move out-of-state before the testing window.

- Students with a medical condition that prevents them from taking the test may be exempted on the basis of a doctor's recommendation and concurrence by KDE. It should be noted that many medical disabilities are accommodated by means of a scribe or a computer. Students with Individual Educational Programs may also receive accommodations and be able to complete the test. Some students do the alternate portfolio.

- Another group that does not take the test are those who have dropped out or graduated before the test date.

- Students who have not been enrolled in Kentucky schools enough days may be exempted from completing the Writing Portfolio.

- Students who are English Language Learners may also be exempted from the test.

As can be seen, Kentucky makes every effort to test every student who possibly may be fairly tested.

What are nonacademic indicators?

Ten percent of a school's accountability score is based on "nonacademic indicators." These are items that are not subject matter oriented but are very important to success. Included in these are the school's percentage of attendance, the percentage who are required to repeat a grade, the percentage who drop out of school during grades 7 through 12, and at the high school the percentage of students who make a successful transition to adult life. This successful transition is demonstrated by such things as becoming employed, joining the armed forces, entering college or a vocational school, and others. Some of these items must be based on data gathered a year earlier than the testing year in order to be complete.

APPENDIX

(THINGS YOU MAY OR MAY NOT WANT TO KNOW)

Testing the Learner Goals

Kentucky has six goals for learners, established by law.

KENTUCKY'S SIX LEARNER GOALS
1. Students shall use basic communication and mathematics skills for purposes and situations they will encounter throughout their lives. 2. Students shall develop their abilities to apply core concepts and principles from mathematics, the sciences, the arts, the humanities, social studies, practical living studies, and vocational studies to what they will encounter throughout their lives. 3. Students shall develop their abilities to become self-sufficient individuals. 4. Students shall develop their abilities to become responsible members of a family, work group, or community, including demonstrating effectiveness in community service. 5. Students shall develop their abilities to think and solve problems in a variety of situations they will encounter in life. 6. Students shall develop their abilities to connect and integrate experiences and new knowledge from all subject matter fields with what they have previously learned and build on past learning experiences to acquire new information through media sources.

Goals three and four are not tested by the KCCT because it is difficult to devise meaningful ways of evaluating these and the evaluation could raise issues of personal privacy.

Distribution across academic expectations

Assessing the quality of the KCCT includes making sure that the test is properly and comprehensively related to the 57 Academic Expectations. At least once per biennium tables are produced that demonstrate the distribution of items across the Academic Expectations. These tables are not included here, but are available in the Technical Reports produced by the Office of Assessment and Accountability and the contractors. These are available upon request.

Distribution across core content

In a fashion similar to distributions across the Academic Expectations, tables of distribution of items with regard to the Core Content are produced annually. These are carefully checked to make sure the test is matching the Blueprint and to provide guidance to the contractor during the building of the six forms in each of the subject areas tested.

Item analysis

To provide evidence of the technical quality of the KCCT a series of item level analyses are performed for each grade and subject area. The following list summarizes some of the analyses conducted.

- Distribution of item scores for open-response items,
- Distribution of corrected item-total correlations for open-response items,
- Distribution of item-theta correlations for open-response items,
- Distribution of N vs. A, P, D biserial correlations for open-response items,
- Distribution of N, A vs. P, D biserial correlations for open-response items,
- Distribution of N, A, P vs. D biserial correlations for open-response items.

A comprehensive overview of the above analyses, which is an enormous amount of statistical data, is available in the various Technical Reports about the KIRIS and CATS systems. These reports are available from the Office of Assessment and Accountability, Kentucky Department of Education.

In 1998 some initial work on differential item functioning (DIF) was begun. The purpose of these studies was to determine if items function differently for subgroups of students, such as males versus females, or African Americans versus Caucasians. While DIF is a requirement for bias to be present, it is not sufficient to indicate bias, which has to be addressed by the Bias Review Committee. A much larger project across all grade levels and in all subject areas was completed in 2001 and a summary of the results is available from the Office of Assessment and Accountability, Kentucky Department of Education.

Scaling

Scaling is the process of making sure that a score on one form of the test means the same thing as a score on a different form. Scaling is also necessary from year to year to make sure that a given score means the same thing year after year. Scaling involves converting raw scores into scale scores (Kentucky uses a scale from 325 to 800) and doing statistical processes that establish the desired comparability. For more information see the Technical Reports to which we previously referred.

Portfolio Audits

Each year 100 schools are selected to participate in a writing portfolio audit. The purpose is to verify that scoring is being done accurately at the school site since the contractors do not score portfolios. Fifty of the schools are selected randomly, although this sample is divided into three groups: elementary, middle and high school. The other fifty schools are selected on the basis of having the greatest amount of change in their portfolios scores, either up or down. The purpose here is to make sure the changes are real. Once a school is selected it sends all its portfolios to the contractor, where professional scorers evaluate the portfolios using the same rubric (scoring guide) that the teachers used. The scores given by the contractor are the final scores. Many schools are right on target. Some schools grade too easily and some schools grade too

hard. In the past, each fall the audited schools would meet with KDE staff during which means of improving the portfolios and means of improving the accuracy of the scoring were discussed.

School Audits

A school audit, or scholastic audit, normally happens when a school does not meet its goal, and scores among the bottom third of the schools that fell below the assistance line. A diverse team of five educators, the make-up of which is established in regulation, visits the school for a week. Each member of the committee has particular responsibilities: data review, budget analysis, classroom visitation, administrative evaluation and other issues. The analysis takes place during four or five days and a lengthy report is written. The guide for the visiting team is called the Standards and Indicators for School Improvement or SISI. This document organizes a school in terms of its success in meeting ten specific standards for excellence. The indicators are specific evidence that the team looks for that indicate the level of functioning of the school. The evaluation for each indicator is on a scale of one to four and provides discussion of specific reasons for the category into which the school is placed. The intention of the audit is to give the school specific guidance concerning its weaknesses that are causing the students to fail to perform better.

HISTORICAL TIMELINE

The following is a brief summary of actions related to Kentucky's system of assessment and accountability. These are actions taken by the Office of Assessment and Accountability and its predecessor, Office of Curriculum, Assessment, and Accountability.

1990 The OAA assisted NAEP in the 1990 8th grade reading assessment.

Technical assistance was elicited for psychometric advice from experts in the field. The National Technical Working Group (later the National Technical Advisory Panel for Assessment and Accountability) was formally established in 1995.

1991 The OAA assisted in the gathering of information for drafting the 75 Academic Expectations (originally referred to as Valued Outcomes).

1992 The OAA assisted with the drafting of the first set of performance standards.

The OAA, in conjunction with contractors, constructed, administered, scored and reported the first KIRIS assessment for the purpose of establishing baselines for the accountability system for schools.

The first teacher groups (later Content Advisory Committees) were formed to participate in writing and selecting the questions for the KIRIS assessment.

In the following years the KIRIS and its successor CATS used a wide variety of assessment types for the purpose of validity, accuracy of assessment and assisting in modifying instruction including multiple-choice (Pre-tested in the spring of 1997 and 1998, and entered “Long-Term” accountability in 1999), open-response, performance events (1993 to 1996), portfolios (Writing all years, mathematics, 1993 to 1996), and on-demand writing.

The OAA supervised through a contractor the administration and scoring of the alternate portfolio, which was included in the accountability system beginning with the 1992-1993 school year.

The OAA assisted NAEP in the 1992 assessment of 4^{th} grade reading and mathematics, and 8^{th} grade mathematics.

Beginning in 1992, item level reporting was begun to improve student motivation. Changes have been made incrementally from 1992 to 2002 to improve the process.

1993 The OAA provided through a contractor the first technical manual with detailed information concerning the assessment.

The OAA provided the first professional development for the District Assessment Coordinators, and provided the first Implementation Guidebook.

The OAA with assistance from the contractors conducted the first audit of Portfolio scores. After scoring accuracy analyses conducted in 1994 and 1995, the audits became a regular feature.

KIRIS Curriculum and Assessment Reports were initiated for purposes of accountability. These later became the KIRIS Performance Reports (1997) and the Kentucky Performance Reports (1999).

1994 The OAA adjusted the assessment process based on the legislative withdrawal of Learner Goals 3 and 4 from assessment, and aided the reformulation of the 75 Valued Outcomes into the 57 Academic Expectations.

The OAA again assisted NAEP in the assessment of 4th grade reading.

The first KIRIS cycle ended with the assignment of rewards and sanctions.

The OAA assisted in the establishing of the first *Content Guidelines.*

The OAA assisted with the production of the portfolio implementation manuals.

1995 The OAA assisted in the study/validation of the 1992 performance standards.

1996 First *Core Content for Assessment* document produced (Revised by Curriculum Division in 1999).

The OAA assisted NAEP in the administration of assessments in 4th grade mathematics and 8th grade mathematics and science.

Assessment Cycle 2 ends with appropriate rewards and sanctions.

1997 The administration of the CTBS/5 Survey Edition began.

1998 The third KIRIS cycle ended with the assignment of rewards and assistance.

The OAA assisted with the NAEP assessments in 4th grade reading, and 8th grade reading and writing.

1999 The CATS Interim Cycle begins.

A two-year multi-step standard setting project was initiated.

The CTBS/5 Survey Edition is included in the Long-Term Accountability index.

The Validation and Research Division was initiated. The OAA has engaged in a continuously expanding program of validation over the decade, assisted by the contractors and focusing primarily upon construct and consequential validity, which led to the creation of the division.

2001 The standards setting process was completed and the new standards for the KCCT were adopted by the Kentucky Board of Education.

Goal lines and assistance lines were recalculated for all schools and growth charts produced based upon the new standards.

The Longitudinal Accountability Pilot Project continued.

A major project to assess Differential Item Functioning (DIF) was initiated to determine if any items were potentially discriminating against a subgroup.